Interspeech 2022
18-22 September 2022, Incheon, Korea

End-to-End Spontaneous Speech Recognition Using Disfluency Labeling

Koharu Horii[1], *Meiko Fukuda*[2], *Kengo Ohta*[3],
Ryota Nishimura[2], *Atsunori Ogawa*[4], *Norihide Kitaoka*[1]

[1]Toyohashi University of Technology
[2]Tokushima University
[3]National Institute of Technology, Anan College
[4]Nippon Telegraph and Telephone Corporation
horii.koharu.fj@tut.jp, kitaoka@tut.jp

Abstract

Spontaneous speech often contains disfluent acoustic features such as fillers and hesitations, which are major causes of errors during automatic speech recognition (ASR). In this paper, we propose a method of "disfluency labeling" to address this problem. Our proposed method replaces disfluent phenomena in the transcription of speech data used for training with two types of labels, filler (＃) and hesitation (@), and trains an end-to-end ASR model using this data, which makes it possible to recognize disfluent acoustic phenomena as recognition targets, like characters. In addition, by removing the disfluency labels that are included in the recognition results, the words that the speaker actually intended to say can be extracted from the disfluent speech. The results of our evaluation experiments show that both the character and sentence error rates were reduced for all of the ASR test sets when disfluency labeling was applied, compared to the baseline method. The proposed method also outperformed other methods intended to reduce disfluency-related errors, even when more disfluent, spontaneous dialog speech was used. This study shows that explicit learning of two disfluent features, fillers and hesitations, is effective in spontaneous speech recognition.

Index Terms: end-to-end speech recognition, spontaneous speech, disfluency, filler, hesitation

1. Introduction

The advent of end-to-end (E2E) learning models has resulted in improvement of the accuracy of automatic speech recognition (ASR) systems. However, spontaneous speech contains fillers such as "uh" and "um" and hesitations such as "I thoug… knew" which traditional ASR systems cannot handle well, making these disfluent acoustic phenomena major causes of misrecognition in ASR [1]. Several approaches have been proposed to lessen this problem.

Researchers have modeled these phenomena using phoneme hidden Markov models (HMMs) [2, 3]. For example, the repetition "yeah, yeah" is modeled using the phonemes "Y AE Y AE" [2]. In [3], hesitations and word fragments are labeled and modeled directly using an HMM and an N-gram language model. When using these HMM-based approaches, however, the model needs to be designed to capture a wide range of acoustic features, due to the constraint of the HMM's structure. In [4], the authors model hidden events such as hesitations using prosodic decision trees. However, due to the integration of acoustic and language models, the models became very complex.

Several studies have also been conducted using E2E ASR approaches. In [5], the authors try to develop a speech recognizer that would generate fluent transcriptions directly from disfluent speech, by implicitly skipping disfluent phenomena. They had hoped that the ASR model could produce fluent, end-to-end transcriptions without explicitly detecting the disfluencies. In [6], hesitations in spontaneous English speech are explicitly tagged by using a hesitation label, which corresponds to any hesitations, and recognize as hesitation labels using an Recurrent neural network (RNN) -Transducer model. At almost the same time with [6], we also proposed an E2E ASR model that can treat hesitations in spontaneous Japanese speech [7] by using the similar labeling scheme to [6].

In this study, we extend our previous proposal [7] and the method proposed in [6]. In addition to hesitations, we also tag fillers using a disfluency label, i.e., we replace fillers and hesitations in transcriptions of a spontaneous Japanese speech corpus by using the disfluency labels '＃' and '@', respectively. The labeled data is then used for training E2E using a joint connectionist temporal classification (CTC)-attention Transformer ASR model. This allows the model to learn the acoustic featuresof the disfluency labels, so that the ASR model can accurately recognize these two types of the disfluencies by replacing them with the labels. By labeling both fillers and hesitations and by removing them from recognition results, we can obtain sentences that are closer to what the speaker originally intended to convey compared to [6, 7], which only label hesitations. The following box shows an example of our disfluency labeling scheme.

Before	:	あのー ひ冷や汗かきながらあーやってます Anō hi hiya-ase kakinagara ā yatte masu
Translation	:	Umm... I'm doing it, ah... while I'm col, cold sweating.
After	:	＃ @冷や汗かきながら＃やってます ＃ @ hiya-ase kakinagara ＃ yatte masu
Translation	:	＃ I'm doing it, ＃ while I'm@ cold sweating.

As shown in the above example, the fillers "あのー (anō)" and "あー (ā)" are replaced by "＃", while the hesitation "ひ冷や汗 (hi hiya-ase)" is replaced by "@冷や汗 (@ hiya-ase)".If we can obtain correctly labeled recognition results, the information that the speaker originally wanted to say can be extracted from their disfluent speech by removing these disfluency labels, as shown below:

10.21437/Interspeech.2022-281

Target : 冷や汗かきながらやってます
Hiya-ase kakinagara yatte masu
Translation : I'm doing it, while I'm cold sweating.

Note that, in this study, we conduct experiments using a Japanese speech corpus, but our proposed method can be applied to ASR of any language. The rest of this paper is organized as follows. In Section 2, we provide the details of our proposed method, which we call "disfluency labeling". In Section 3, we describe our ASR experiments and report our results. In Section 4, we discuss the results of this study. Finally, we conclude this paper in Section 5.

2. Disfluency labeling

Spontaneous speech often includes disfluent acoustic phenomena such as fillers and hesitations. These types of disfluencies are major causes of misrecognition when performing ASR. Since the sentence that the speaker intended to say does not contain these phenomena, we need to remove them from the recognition results to interpret the actual target sentence. In this section, we describe in detail a method of labeling disfluent acoustic phenomena in speech so that they can be treated as a single recognition target, like characters. We call this method "disfluency labeling".

Examples of disfluent acoustic phenomena are shown in Table 1. The tags shown in this table are based on the tagging criteria [8] of the Corpus of Spontaneous Japanese (CSJ) [9], which is the spontaneous speech data set used in our experiment, as described in Section 3.1. In this study, according to the labeling rules of CSJ, we treat "fillers", "interjections"[1], and "responses"[1] as fillers, and "stammering" and "restatement" as hesitations. The utterances which are designated as disfluencies are those enclosed within brackets. Some of the utterances which are provided as examples of fillers in the table are similar to Japanese adverbs and collocations, but they can be identified as fillers based on the context. "interjections" and "responses" are not, strictly speaking, fillers, but since they are often indistinguishable, they are treated here as fillers. Word fragments such as those shown in the "stammering" section of the table are treated as hesitations. Repetition or rephrasing of entire words that make sense as words are not treated as hesitations. Even if an utterance is a word fragment, it is not considered to be hesitation if it differs from the first syllable of the following word and is uttered without rephrasing. In the "restatement" section of Table 1, the words in brackets have the same meanings as the following words, but the speaker has decided to use a different word, thus the first word is classified as a hesitation. If there are several consecutive fillers or hesitations, they are not grouped together but are treated as individual disfluencies. Our proposed method is to replace the utterances enclosed within brackets with disfluency labels, thus we replace fillers and hesitations with the labels '＃' and '@', respectively.

We expect the model to learn these labels along with the other characters when using this method. This should increase recognition accuracy, because the model should be able to learn the acoustic features of fillers and hesitations, then seamlessly recognize them, preventing misrecognition when they are encountered. Unlike implicit removal methods, such as the one proposed in [5], the use of labels provides the model with information on the locations of the disfluencies, which is expected to result in better recognition accuracy. Our method is also able to obtain the speaker's intended sentence by removing the labeled disfluent utterances from the recognition results. In [6] and [7], only hesitation labels were used, but our method also labels fillers in order to obtain sentences that are closer to what the speaker originally intended to convey.

[1]Since the CSJ's policies are followed, these are treated as fillers, but it is debatable whether they should be classified as fillers.

3. Evaluation experiments

3.1. Corpus

In order to evaluate our proposed method, we conducted evaluation experiments using the CSJ, which contains 661 hours of spontaneous Japanese speech, as well as transcriptions and various additional information for experiments. It contains approximately 7 million words, and consists of academic presentation speech (APS), simulated public speech (SPS), readings, dialogs, and so on [9]. APS and SPS account for 90% of the data. It also contains a test dataset consisting of speech from 30 speakers. The test dataset is divided into three groups of ten speakers each. Hereafter, we refer to them as Eval1, Eval2, and

Table 1: *Examples of disfluency labeling based on CSJ labeling criteria.*

Label	Disfluency type	Example		Labeled example
＃	Filler	[まー]	[mā]	＃
		[あのですね]	[ano desune]	＃
		[えっと] [あの] [んー]	[etto] [ano] [nn]	＃ ＃ ＃
	Interjection	[うわっ]	[uwa']	＃
		[えー]	[ē]	＃
		[おー]	[ō]	＃
	Response	[はい]	[hai]	＃
		[ええ]	[ē]	＃
		[いいえ]	[ī̄e]	＃
@	Stammering	[喋っ] 喋った	[shabe'] shabetta	@ 喋った
		[こ] 来ない	[ko] konai	@ 来ない
		[さ] [最] 最大の	[sa] [sai] saidaino	@ @ 最大の
	Restatement	[テビス] テニス を する	[tebisu] tenisu o suru	@ テニス を する
		明日 [ですので] ですから	ashita [desunode] desukara	明日 @ ですから
		千六百[三十]三年	sen roppyaku [sanjū] san nen	千六百 @ 三年

Table 2: *Expected raw output and reference sentences for each model.*

Model	Expected raw output	Reference sentences
Baseline	えー いじ あっ以上で大体発表終わります	えー いじ あっ以上で大体発表終わります
HL [6, 7]	えー @ あっ以上で大体発表終わります	えーあっ以上で大体発表終わります
DR [5]	以上で大体発表終わります	以上で大体発表終わります
Proposed	# @ #以上で大体発表終わります	以上で大体発表終わります

Eval3. Eval1 and Eval2 contained APS while Eval3 contained SPS. The number of utterances in the test sets were 1,272 for Eval1, 1,292 for Eval2, and 1,385 for Eval3. The CSJ also contains 58 dialogs, with separate channels for each speaker. The dialog speech contained 2,363 utterances from interviews with presenters of APS and SPS, task-oriented dialog, and free dialog. Since this dialog speech is more spontaneous than the APS and SPS speech, in addition to the three test sets described above, we also used CSJ dialog speech for evaluation. The model was trained using the APS and SPS training sets, which does not overlapped with the test sets or the dialogs.The training data consisted of 413,377 utterances, while the development data consisted of 4,000 utterances. One type of additional information included in the CSJ are tags identifying fillers ('F') and hesitations ('D' and 'D2') [8]. When performing disfluency labeling in this study, we converted these tags and the targeted utterances into disfluency labels ('#' and '@', respectively).

3.2. Experimental settings

Experiments were conducted using the E2E speech processing toolkit ESPnet2 [10] (ver. 0.9.9) on a computer equipped with one NVIDIA GeForce RTX 3090 GPU.The baseline model was a joint CTC-attention Transformer ASR model of ESPnet2 trained using CSJ without disfluency labeling. Training was repeated for 20 epochs. To maximize speech recognition performance, we saved the model parameters of the 10-best epochs according to the validation sets, and averaged them at the end of training. To improve accuracy, SpecAugment [11] and speed perturbation (SP) [12] were also applied. The SP speed coefficients were set at 0.9, 1.0, and 1.1, and data expansion was performed by changing the speech speed. The ESPnet2 default settings were used for the hyperparameters for training and recognition. In this study, speech recognition was performed using the ASR model without a language model.[2]

We trained the ASR model for the proposed method using data labeled with the proposed disfluency labeling method, in which fillers were replaced by '#' and hesitations by '@'. ASR results were compared to the baseline, the disfluency removal (DR) model [5], which was trained by removing the disfluencies, and the hesitation labeling model (HL) [6, 7], which was trained by labeling only the hesitations. All of the models were evaluated using the same CSJ test sets. In addition, we also evaluated recognition performance using the dialog speech dataset, which contains utterances with a higher degree of spontaneity. The reference sentences for our proposed method, DR, and HL were obtained by labeling them in the same manner as the training data for each model, and removing the labels from the recognition results. The expected recognition results and reference sentences for each model are shown in Table 2. Accuracy was measured using the character error rate (CER) and sentence error rate (SER).

[2]We conducted pilot experiments using a language model with shallow fusion, but only negligible differences in accuracy were obtained.

Table 3: *Experimental results (regular CSJ test sets).*

Model	Data	CER [%]	SER [%]
Baseline	Eval1	5.4	53.8
	Eval2	3.8	50.3
	Eval3	4.5	36.5
HL [6, 7]	Eval1	5.0	49.8
	Eval2	3.5	47.8
	Eval3	4.1	34.2
DR [5]	Eval1	4.6	43.2
	Eval2	**3.2**	39.6
	Eval3	3.9	30.2
Proposed	Eval1	**4.5**	**41.6**
	Eval2	**3.2**	**37.5**
	Eval3	**3.8**	**28.4**

Table 4: *Experimental results (CSJ dialog test set).*

Model	CER [%]	SER [%]
Baseline	16.0	62.8
HL [6, 7]	15.5	60.9
DR [5]	12.9	49.9
Proposed	**10.3**	**32.8**

Table 5: *Breakdown of CERs for CSJ dialog test set.*

Model	Error rates [%]		
	Substitution	Deletion	Insertion
Baseline	7.6	6.2	**2.1**
HL [6, 7]	6.7	6.4	2.4
DR [5]	**5.2**	2.4	5.3
Proposed	**5.2**	**2.3**	2.8

3.3. Experimental results

Our experimental results using the CSJ test sets (Eval 1, 2 and 3) are shown in Table 3. Compared to the baseline and other models, the proposed model achieved better speech recognition accuracy for all of the regular evaluation data, outperforming the HL method, which labeled only hesitations[6, 7]. The reduction in SER between the baseline and HL methods was 6.3% on average, while the reduction between baseline and the proposed methods was 23.5% on average, indicating that filler labeling is useful in improving accuracy. The DR method targets both

fillers and hesitations, but our proposed model, which learns the location and type of disfluent phenomena, achieved higher accuracy, which suggests the effectiveness of explicit learning of this information. The proposed method achieved a CER of less than 4.5% and an SER of less than 42% for all of the evaluation data. The error rate was especially low for the Eval2 data, with a CER of 3.2%, and for the Eval3 data with a SER of 28.4%. This may be the upper boundary for ASR accuracy, considering that the manually transcribed reference sentences contained some errors and notational distortions.

Results for the dialog speech recognition task are shown in Table 4. The proposed model was the most accurate in this task as well. Note, however, that, due to the nature of dialog speech, there were many utterances consisting of only interjections or responses, which were treated as fillers. The baseline results also show that the dialog speech was more difficult to recognize than the APS and SPS data due to its higher spontaneity. Our proposed model achieved significant improvement over the other methods, achieving a CER of 10.3% and a SER of 32.8%. The SER was about half that of the baseline method. A breakdown of the CERs for each method is shown in Table 5, showing that the HL model had the highest deletion error rate. The DR model shows closer performance to that of the proposed model for the regular test sets, but for the dialog speech the proposed model achieved a 20.2% lower CER and a 34.3% lower SER than the DR model, which are significant differences. The DR model had the highest insertion error rate, suggesting that this model does not adequately address disfluency in spontaneous dialog tasks. These results indicate that the proposed method is more effective for speech recognition with more spontaneous speech than the other models. Examples of recognition results for the proposed method are shown below:

Pronunciation	:	ん えー ちょ聴覚フィルターに N ē cho chōkaku-firutā ni 対応するえーとー taiō suru ētō
Reference	:	# # @聴覚フィルターに対応する# # # @ chōkaku-firutā ni taiō suru#
Result	:	# # @聴覚フィルターに対応する# # # @ chōkaku-firutā ni taiō suru#
Target	:	聴覚フィルターに対応する chōkaku-firutā ni taiō suru
Translation	:	Corresponds to an auditory filter.

Pronunciation	:	うーん嬉しいって言うよりし信じられない Ūn ureshii tte iu yori shi shinzirarenai
Reference	:	#嬉しいって言うより@信じられない # ureshii tte iu yori @ shinzirarenai
Result	:	#嬉しいって言うより@信じられない # ureshii tte iu yori @ shinzirarenai
Target	:	嬉しいって言うより信じられない ureshii tte iu yori shinzirarenai
Translation	:	I' m more incredulous than happy.

Since the labels appear in the same positions as in the reference sentences, this suggests that the model has correctly learned the acoustic features of the labels, and that labeling is being performed correctly. Furthermore, by removing the labels from the recognition results, the speakers' target sentences could be obtained from their disfluent, spontaneous utterances.

4. Discussion

In this study, disfluency labeling was used to allow the disfluent acoustic phenomena known as fillers and hesitations in spontaneous speech to be represented by the recognition targets '#' and '@', respectively, and treated as with other characters. Based on the experimental results of the present study, we found that, when performing ASR with spontaneous speech, it was more effective to explicitly learn both types of disfluencies, fillers and hesitations. In order to further improve speech recognition accuracy, it would worthwhile to investigate an approach that is able to analyze the contexts in which spoken language tends to become disfluent, as well as words which are prone to mispronunciation. Pseudo-labels could be assigned to a huge amount of written language data, which would then be used to train a language model how to use these labels. The language model could then be combined with an ASR model. However, it is necessary to take into account that spoken and written language differ in terms of vocabulary and sentence structure.

In this study, using the CSJ policy that short responses should be labeled as fillers, as shown in the Table 1, our model accurately removed them, but some of these responses had an implied "yes" or "no" meaning, which should not have been removed in order to preserve the nature of the conversation. In addition, some short Japanese responses, such as "un", are acoustically similar to fillers, so they tended to be recognized as a type of filler in our evaluation experiment, even though these responses only appeared in the dialog speech, and thus were not trained. To tackle this problem, an approach which considers a longer context, which may include semantic information, could be used to judge whether or not an utterance is a filler.

5. Conclusions

In this study, we proposed a method of disfluency labeling to allow disfluent acoustic phenomena, such as fillers and hesitations in spontaneous speech, to be treated as single recognition targets, like characters. Compared to a baseline and previously proposed methods used to reduce the impact of disfluencies on speech recognition accuracy, our proposed method showed the best recognition accuracy, both in a task involving presentation speech and a task involving spontaneous dialog speech. These result shows that when recognizing spontaneous speech, it is effective for models to explicitly learn two kind of disfluencies, fillers and hesitations. In addition, by removing the disfluency labels after processing, the sentences that the speaker actually intended to say can be easily extracted. Although the proposed method achieved recognition accuracy considered to be close to the upper limits of ASR performance, a contextual approach should also be explored to further improve accuracy in the future.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP19H01125, JP19K04311, and JP21K13641.

7. References

[1] S. Goldwater, D. Jurafsky, and C. Manning, "Which Words Are Hard to Recognize? Prosodic, Lexical, and Disfluency Factors that Increase ASR Error Rates." 01 2008, pp. 380–388.

[2] V. Rangarajan and S. Narayanan, "Analysis of disfluent repetitions in spontaneous speech recognition," in *2006 14th European Signal Processing Conference (EUSIPCO)*, 2006, pp. 1–5.

[3] R. Rose and G. Riccardi, "Modeling disfluency and background events in ASR for a natural language understanding task," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, vol. 1, 1999, pp. 341–344.

[4] A. Stolcke, E. Shriberg, D. Z. Hakkani-Tür, and G. Tür, "Modeling the prosody of hidden events for improved word recognition," in *EUROSPEECH*, 1999.

[5] P. J. Lou and M. Johnson, "End-to-End Speech Recognition and Disfluency Removal," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 01 2020, pp. 2051–2061.

[6] V. Mendelev, T. Raissi, G. Camporese, and M. Giollo, "Improved Robustness to Disfluencies in RNN-Transducer Based Speech Recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6878–6882.

[7] K. Horii, M. Fukuda, K. Ohta, R. Nishimura, A. Ogawa, and N. Kitaoka, "End-to-End Spontaneous Speech Recognition Using Hesitation Labeling," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 1077–1081.

[8] The National Institute for Japanese Language, "Construction of the Corpus of Spontaneous Japanese," in *The National Language Research Institute Research Report No. 124*, 2006.

[9] K. Maekawa, "Corpus of Spontaneous Japanese : its design and evaluation," *Proceedings of The ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, pp. 7–12, 2003.

[10] S. Watanabe, F. Boyer, X. Chang, P. Guo, T. Hayashi, Y. Higuchi, T. Hori, W.-C. Huang, H. Inaguma, N. Kamo, S. Karita, C. Li, J. Shi, A. S. Subramanian, and W. Zhang, "The 2020 ESPnet Update: New Features, Broadened Applications, Performance Improvements, and Future Plans," in *2021 IEEE Data Science and Learning Workshop (DSLW)*, 2021, pp. 1–6.

[11] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[12] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech 2015*, 2015, pp. 3586–3589.